## SUMMARY

# FUNCTIONAL PROTEOMICS
# IN ENVIRONMENTAL HEALTH SCIENCE

*Sponsored by*

**The Southwest Environmental Health Sciences Center,
The Arizona Cancer Center**

*and the*

**National Institute of Environmental Heath Sciences
National Center for Toxicogenomics**

**January 24, 2001**

# Summary of Workshop on
# Functional Proteomics in Environmental Health Science

January 24, 2001
Peter Kiewit Auditorium, Arizona Cancer Center
The University of Arizona, Tucson AZ

## Introduction

The National Institute of Environmental Health Sciences (NIEHS) established the National Center for Toxicogenomics (NCT) in June 2000. Toxicogenomics is an emerging scientific field that combines studies of genetics, genome-wide mRNA expression, cell and tissue-wide protein expression, and bioinformatics to understand the roles of gene-environment interactions in disease. The NCT was created to facilitate application of toxicogenomics to improve human health.

The goals of the NCT are: 1) to develop and apply gene expression and proteomics technology to study the biological effects of chemicals and drugs; 2) to support intramural and extramural research to define the effects of environmental agents on gene expression; and 3) to develop a national reference and relational database on "Chemical Effects and Biological Systems" (CEBS) that will serve as a resource in the fields of toxicology and environmental health.

To help define the path towards fulfilling its goals, a series of 3 workshops are being held through which the research community can learn about the NCT and provide input to the NCT. The first workshop in this series entitled "Functional Genomics and Environmental Health" was held at Massachusetts Institute of Technology in Cambridge, Massachusetts on December 11, 2000 (report of this meeting will be available on the internet.) This meeting summary reports on the second workshop entitled "Functional Proteomics in Environmental Health Science," which was held on January 24, 2001 at the University of Arizona in Tucson, Arizona and was co-organized by Dan Liebler (University of Arizona, Tucson) and Kenneth Tomer (NIEHS). The workshop was co-sponsored by NIEHS, the Southwest Environmental Health Sciences Center at the University of Arizona and the Arizona Cancer Center.

Dan Liebler (Director, Southwest Environmental Health Sciences Center) and Ray Nagle (Deputy Director, Arizona Cancer Center) welcomed the workshop participants and expressed their optimism about the future of toxicogenomics and proteomics. Ray Tennant (Director, NCT, NIEHS) provided an introduction to the morning session. He stated that the goal of the workshop was to provide information needed to develop a solid foundation for the NCT. Toxicogenomics is providing new avenues to solve previously intractable problems in toxicology. For example, toxicogenomics offers new means to identify hazardous compounds and individuals exposed to those compounds, and means to predict and/or prevent disease. To achieve these objectives, NCT will need to develop a robust database. NCT intends to facilitate development of this database and its associated essential bioinformatics tools, and stimulate hypothesis-driven research that will continue to enrich this database.

The NCT will play a major role in forming partnerships between diverse sectors of the research community.  It will coordinate and form research consortia and engage a contractor to oversee development and maintenance of the CEBS database.  One goal of this workshop is to provide a forum for input on the future structure and operation of the NCT.  Thus, Tennant solicited comments from the audience during the open discussion session of the workshop.

**Proteomic Profiling with 2-Dimensional Gels**

Julio Celis (University of Aarhus, Denmark) described proteomic profiling with 2-dimensional (2D) gel electrophoresis to identify biomarkers of bladder cancer.  His presentation focused on approaches to identify premalignant lesions in tissue samples and on protein biomarkers abundant in the urine of bladder cancer patients.

Celis pointed out that the proteome of a single type of cell varies according to its environment and its growth state.  Thus, proteomics examines not one proteome, but many different proteomes.  The field of proteomics has made great progress in recent years because of the coupling of mass spectrometry with 2D gel electrophoresis.  These recent technical advances have been important, because proteins are intrinsically harder to work with than mRNA or DNA.  Proteomics has many applications; it is used to separate, identify and quantify proteins, and to analyze protein function.  In addition, proteomics is a powerful method to identify and quantify protein modification (i.e., phosphorylation, glycosylation).

The technology of high resolution 2D protein gel electrophoresis has been in use for approximately 25 years.  It is an excellent method to separate and identify proteins, and allows information to be stored directly in the gel itself.  Using fractionated $^{35}$S-labeled protein from a mammalian cell, approximately 3,500 proteins can be detected on a single gel, with up to 6000 mammalian proteins detectable by this technology.  This is estimated to be approximately 30% of the total mammalian proteome.  Celis (and other workshop participants) emphasized that a major challenge in working with the proteome is to detect protein species over a very large dynamic range.  Thus, much work in proteomics has been qualitative instead of quantitative.  Specialized approaches are used for low abundance proteins including antibody detection and fluorescent dye labeling.  For example, in Celis' hands, an anti-Ras antibody detects Ras at a level of 20,000 molecules per cell.  Fractionation can be used to increase detection limit by enriching for a protein of interest, but this approach is not applicable if high throughput is needed.

Celis described experiments carried out in the 1980s, when scientists were first able to identify a single protein species on a 2D gel.  His studies of transformed NIH3T3 cells and of cells from patients with lupus erythematosus identified a 49 kDa protein he called "cyclin" present only in proliferating cells.  Later studies showed that this species is the now well characterized replication factor proliferating cell nuclear antigen (PCNA).

Proteomics involves cataloguing each protein species on a gel.  This requires methods for master numbering, gel matching, quantitation, annotation, database management and database linking.  Annotation of each protein species is critical and should include information such as name, localization, protein sequence, function, expression, gene, etc.  It will be useful to have special databases for different cell types, tissues and cellular fluids.  In some cases, there are limits on the information that can be placed in a public database, because of clinical implications and restrictions on proprietary information.  It is and will continue to be difficult to compare information in protein databases and cDNA expression databases.

Squamous cell carcinoma (SCC) is a common disease striking many patients with abundant tumors in the bladder and other tissues. Celis used 2D gels to characterize the cellular progression from transitional epithelium, to keratinocyte metaplasia, to premalignant lesion, and finally to SCC in patients with bladder tumors. Deep biopsies were taken that include normal and tumor cells, 2D gels were run and blotted, and samples were probed with high quality antibodies for keratins. The results identified several keratins (K7, K13, K14, K19 and K29) that are differentially expressed in normal or tumor cells at different stages. These results were confirmed with specific sectioning and immunohistochemistry, demonstrating that K13 and K19 are differentially expressed in normal cells, and K14 is differentially expressed in adjacent tumor cells in the same biopsy sample. These results and the results of additional studies were used to develop criteria by which cells can be characterized as normal or tumor cells of different stages based on which keratin species they express. Celis also used 2D gels to identify a protein specifically expressed in SCC tumor cells called psoriasin that is secreted into urine. Analysis of the urine of SCC patients identified an abundant signal for this protein, which may be suitable as a biomarker for early stages of SCC.

Proteomics is a powerful technology which can help researchers understand the coordination and regulation of many cellular processes. It is also valuable to assess gene knockouts and as a tool for molecular pathology. To develop its full potential, proteomics will need to develop more robust methods for quantification and continue to improve bioinformatics support.

**Mining Proteomes with Liquid-Chromatography Tandem Mass Spectrometry**

John Yates (Scripps Research Institute) summarized state-of-the-art proteomics using liquid chromatography tandem mass spectrometry (LC-MS-MS). Yates pointed out that opportunities in the field of proteomics are expanding as genome sequences, including the human genome, are being completed rapidly. However, gene expression analysis is less demanding than protein expression analysis (as pointed out by Celis) because of the dynamic range problem inherent in study of protein expression. Proteomics has a significant advantage in that it can readily identify structural or translational modifications of proteins, which are important regulatory mechanisms in many cellular processes.

Yates provided an overview of LC-MS-MS techniques. There are two major mass spectrometry techniques called matrix assisted laser desorption ionization-time of flight (MALDI-TOF) and Tandem MS (MS-MS). In MALDI-TOF, ionization is followed directly by mass analysis with one-to-one data correspondence. In Tandem MS, ionization and separation occur in the first step and in the second step, selected ions are fragmented in a collision chamber to yield amino acid sequence data. Tandem MS provides a high accuracy of mass analysis, which can lead to unambiguous protein identification based on the mass of peptide fragments and amino acid sequence. Protein identification can be made by querying protein and nucleic acid databases with the determined amino acid sequences.

Protein mass mapping is generally carried out by extensive trypsin digestion of a protein of interest followed by high resolution mass spectrometry, which yields a unique protein fingerprint. This technique is best used with a homogeneous protein. However, LC-MS-MS of peptide mixtures is feasible, with appropriate methods of data analysis. Yates and colleagues developed multidimensional LC and an algorithm called SEQUEST to solve this data analysis problem. Multidimensional LC uses a dual capacity column that carries out ion exchange separation and reverse phase separations in a sequential manner. The protein mixtures are digested, which increases the sample complexity, but also increases sample solubility and coverage of the analysis. Sample loss is minimized because the protein digest acts as a carrier and reduces adsorption of low level analytes. SEQUEST is a powerful database searching algorithm that achieves impressive output speed. It can be used with a bank of linked CPUs, obviating the need for higher capacity computer hardware.

Yates successfully applied shotgun analysis of peptide mixtures to study the yeast ribosome, proteosome and to the total yeast proteome. Using a total yeast cell lysate, 2000 proteins were detected of the theoretically possible 6200 species. The detected species covered a dynamic range of approximately 1000/1, and a range of 9000/1 was detected in a protein spiking experiment. The yeast membrane proteins represented in this data set were also examined. Approximately half of the 300 known yeast membrane proteins with more than 3 transmembrane domains were detected. Yates also described methods for quantification (i.e., $^{15}$N labeling, isotopically labeled covalent modifiers, and isotopically labeled covalent affinity) and for studies of protein phosphorylation in human cells.

The shotgun peptide analysis approach described by Yates utilizes powerful LC-MS-MS technology combined with an efficient searching algorithm. It is a relatively low cost and moderate throughput technique that can be used to analyze many types of highly complex protein samples. Wide application of this approach is expected to help resolve many important questions about the proteome and other questions in biology.

**Proteomics to Characterize Targets of Environmental Chemicals**

Dan Liebler described how proteomics can be used to understand the biological effects of environmental chemicals. Much effort is focused on identifying DNA adducts that result from reactive metabolites following environmental exposure. However, protein adducts are formed at a significant rate and these adducts also result in biological effects that may have toxicological importance. Until recently, these adducts were not well characterized and specific adduct sites on a specific protein were not readily identified. Liebler has developed an LC-MS-MS technique specifically designed for study of protein adducts.

Data-dependent scanning is an instrument control technique that automatically acquires MS-MS spectra of large numbers of peptides in a mixture. In analyses to detect peptide adducts, one does not necessarily know which proteins are adducted, so the acquisition of MS-MS data for as many proteins as possible increases the likelihood of adduct detection. However, data-dependent scanning generates thousands of MS-MS spectra per LC-MS analysis. To help detect MS-MS scans corresponding specifically to adducted peptides, the Liebler group designed a data reduction algorithm called SALSA (Scoring ALgorithm for Spectral Analysis).

Modified peptides form diagnostic products during fragmentation, and these products can include neutral loss, charge loss, ion pairs or product ions. SALSA is optimized to identify these diagnostic products. Liebler has applied SALSA to identify benzoquinone adducts, methylated tyrosine adducts in methyl methanesulfonate-treated BSA and carboxymethylated cysteine residues in phospholipase 2 and other proteins following exposure to dichloroethylene. The experiments demonstrate that low abundance modifications can be detected, although sensitivity may still be limiting in some cases. The SALSA algorithm can be used successfully when the protein modifications are not known, as in cases of unknown exposure to more than one environmental agent. These studies show that proteomics can make a significant contribution to understanding toxic effects mediated through modification of proteins. LC-MS-MS combined with the SALSA algorithm is an important tool for obtaining information on environmentally-induced protein adducts that may have significant biological consequences.

**Application of Proteomics of Eukaryotic Signal Transduction and Tumorigenesis**

Katheryn Resing (University of Colorado, Boulder) described studies of eukaryotic cells with 2D gel electrophoresis and MALDI-TOF/MS. Resing's procedure involved separation by 2D gel electrophoresis, gel excision, in-gel digestion, MALDI-TOF/MS analysis, and database searching with the ProFound™ algorithm. Resing indicated the importance of checking 2D gel data for reproducibility. Error graphs were prepared which indicate higher variability for low intensity spots; an error of less than 30% is considered acceptable. Using a 325 micron i.d. column and an electrospray ionization interface, the working sensitivity of their method was 350 fmol. This sensitivity improved to 30 fmol using a 75 micron i.d. column, which is more difficult to prepare and run. For low abundance species, it was necessary to pool samples from several gels to achieve these results. For example, in one experiment it was possible to detect MAP kinase kinase 2 using spots from 31 gels; the protein is estimated to be present at a level of 5,000-8,000 molecules per cell.

Resing compared the proteome of melanoma cell lines that have different metastatic potential. Differences in protein expression were associated with different cell lines. Some changes correlated with changes in other premalignant cells (i.e., WM35 control cells with oncogenic Ras). Some of the proteins identified are previously known markers for melanoma, and other proteins are new species. Resing suggested that the stages of melanoma progression from nonmetastatic to increasingly metastatic cells could be characterized by this approach.

Gene expression studies were also carried out with human K562 leukemia cells treated with phorbol ester (TPA) to activate protein kinase C (PKC). TPA-induced proteins were identified and confirmed using a PKC inhibitor. Cluster analysis of the data revealed 91 spots corresponding to 41 different induced proteins. Many of the induced proteins were expressed in both a modified and unmodified form. One protein of interest identified in this study is a human homologue of RAD23 (HR23). HR23 is known to interact with the DNA repair protein XPC and with the proteosome.

**Proteomics and Functional Genomics Applied to Drug Discovery**

Stanley Hefta (Bristol Myers Squibb) described the efforts of the Department of Applied Genomics at Bristol Myers Squibb (BMS) to apply genomics and proteomics to drug discovery. Research and development of pharmaceuticals is a long-term and expensive process. In the past, there was often a limited amount of information available on proteins, genes and functions that were relevant to a specific research goal; thus, drug development began with the selection of a limited number of targets. The limited number of targets, and the lack of information about their function, led to a bottleneck when it became necessary to develop a high throughput screening method. Thus, the research and development phase of drug discovery has often required 8-10 years at a cost of over $500 million per drug. And yet, many products fail at late stages in the development program. Genomics and proteomics have changed this picture significantly, most notably by increasing the number of available drug targets by two or more orders of magnitude, by providing essential information on protein and gene function, by speeding up the drug development process, and by saving research time and money.

BMS has developed a number of approaches to use genomic and proteomic-based technology in drug discovery. They employ a wide range of techniques including gene mining, gene sequencing, genotyping, genetics, model organisms, cDNA microarray, robotics, 2D protein gels, mass spectrometry, protein databases and bioinformatics. Many of these technologies are employed through internal and external alliances inside and outside of BMS. BMS participates in collaborative efforts with Millennium Pharmaceuticals, the SNP Consortium, 3-Dimensional Pharmaceuticals, Exelixis Pharmaceuticals, Lexicon Genetics, Affymetrix, Molecular Dynamics, the Whitehead Institute and others.

Hefta described a few examples to demonstrate how BMS uses genomics and proteomics in drug discovery. BMS researchers had developed a candidate drug for prevention of atherosclerosis called BMS433443. This drug was not well understood mechanistically, but it was being studied because the drug induces apoA1, a major structural protein in high density lipoprotein (HDL) and inhibits apoB. The drug is expected to enhance the level of HDL and thus combat atherosclerosis. BMS433443 and a related inactive compound BMS432257 were tested for effects on gene expression using cDNA microarray. Gene expression was assayed with an array of 15000 human genes in the presence and absence drug. The drug-treated cells showed many changes in gene expression, and the pattern of induction suggested strongly that BMS433443 was a potent mitogen. These results were clear enough to warrant termination of all research on BMS433443. While this is not a "success story" of drug discovery, the outcome is a large savings in research cost to BMS, because wasted research effort could be avoided.

BMS uses model organisms in drug discovery which have conserved metabolic pathways, short life span, and small genomes. Researchers were studying the processing of amyloid precursor protein (APP), which plays an important role in the etiology of Alzheimer's disease. They used the model organism *Caenorhabditis elegans* to characterize an inhibitor that may prevent proteolytic processing of APP, because it is thought that such inhibitors may help prevent or moderate symptoms of Alzheimer's disease. The protein target for the inhibitor was suspected to be presenilin, which is also linked to Alzheimer's disease and is involved in proteolytic events in the Notch signal transduction pathway. It was shown that mutants of *C. elegans* deficient in presenilin 1 or presenilin 2 have a strong Notch phenotype and that mutants with a knockout in both genes are nonviable. If wild type *C. elegans* are treated with the putative presenilin inhibitor, the animals show symptoms that mimic a Notch phenotype. However, when *C. elegans* mutants deficient in presenilin1 are treated with the inhibitor, it does not change the mutant phenotype. Further, when the Notch pathway is preactivated, treatment with the inhibitor restores a normal phenotype. Together, these results indicate strongly that the inhibitor targets presenilin in *C. elegans* and further study of the compound for treatment of Alzheimer's disease is warranted.

BMS is also using proteomics techniques including 2D gel electrophoresis, MALDI-TOF and LC-MS-MS. Analyses are carried out with SEQUEST and a proprietary BMS data management software. These techniques have been used in a integrated genomics/proteomics approach to discover novel antibiotics. For example, a set of bacterial mutants was created with deficiencies in metabolic pathways. The protein expression profiles of these mutants were characterized by cDNA microarray and MALDI-TOF. Then antibiotic compounds were screened for their ability to alter bacterial gene expression and the results compared to the expression patterns of mutant cells. Whole proteome analysis is carried out for 3000 *Escherichia coli* proteins in a single step LC-MS-MS run with multidimensional chromatography. Approximately 115 antibiotics have been found that alter major metabolic pathways in *E. coli* which have potential as good targets for drug discovery.

In summary, BMS is building an integrated approach to drug discovery using resources from biology, chemistry, genomics, proteomics and bioinformatics. Internal and external alliances are key components of drug discovery at BMS. It will be critical to develop the capacity and tools for information management, as genomic and proteomic data accumulate rapidly and as technology continues to change and improve at a rapid pace.

## Roundtable Discussion*

The afternoon session of the workshop was a roundtable discussion designed to promote free exchange of information between all participants. Discussants were presented with the following list of questions:

- How can proteome analysis be used for examining the influence of environmental exposure and disease processes on cell function?
- What are the prospects for proteomics in developing clinical and environmental biomarkers for novel gene discovery?
- What new and emerging techniques in proteomics deserve attention and what biological questions are they best suited to address?
- What methodologies and techniques are most effective for proteomics analysis.
- What steps should NIEHS take to enhance the capacity and impact of proteomics?
- How can proteomics address protein-protein interactions?
- How do we integrate molecular structure with protein function?
- What efforts to ensure uniform standards and practices in proteomic analysis are needed?
- How can proteomic and microarray data be interfaced?
- What bioinformatics needs does proteomics create?

The discussion touched on many of these issues. In general, the discussants did not arrive at definitive answers to specific questions, but there were several points of consensus or strongly expressed opinion. These discussion points are summarized below, grouped according to topic.

**Areas of focus:**

It was suggested that it is important for NCT to establish research focus areas. These focus areas should define important biological questions that can be addressed by toxicogenomics research. NCT should emphasize hypothesis-driven research by high quality researchers. Several discussants cautioned against overemphasis on technical issues in toxicogenomics.

The NCT initiative was contrasted with the Human Genome Project. The Human Genome Project was a very large and ambitious undertaking, but the goal was extremely well defined. In contrast, expression profiling and proteomics are exceedingly complicated, much less well defined, and it will be a much more difficult enterprise. Therefore, NCT should select focus areas for toxicogenomic studies in environmental health. Such focus areas are needed to bring ideas and people together to work on a particular subject area, and to enhance efficient progress towards research goals. However, such defined research problems should be flexible enough to encourage maximum creativity by prospective grantees.

It was also suggested that NCT consider focusing all of its effort on one very large problem in toxicogenomics that could not be addressed by a smaller initiative. If successful, NCT would make a large impact by taking this approach.

**Integrated approach: expression profiling and proteomics**

There was clear consensus that NCT should exploit expression profiling and proteomics as appropriate and as needed, using one technology to complement the other. It is also important to acknowledge that gene expression and protein expression data are distinct; one data set can not substitute for the other, because the results are not always directly correlated with one another.

Several discussants emphasized that these technologies offer a way to get leads and initial results; they are often most useful to provide clues to valuable new experimental avenues, but not necessarily for detailed mechanistic studies. Validation studies and further experimentation are usually needed to extend and confirm the results of microarray and proteomics experiments.

Proteomics has a distinct advantage for some studies, especially in areas involving modified proteins. Proteomics also has limitations, because comprehensive information is only available for a few organisms. Microarray is much higher throughput than proteomics, and therefore it is more appropriate for broad-based exploratory experiments. Regardless of the technology used, it will be essential that NCT consider data from model organisms, human cells and human clinical studies.

**Proteomics technology**

The detection limit of proteomics techniques was discussed in regard to the applicability of the technology to study low abundance proteins such as transcription factors. The exact detection limit was not generally agreed upon, but in some cases proteins have been detected at the unusually low level of approximately 100 molecules per cell. More commonly, the detection limit is several thousand molecules per cell. The detection limit is lowered by special staining techniques or by methods that enrich for the proteins of interest. It was agreed that transcription factors can be studied using proteomics technology, and that special approaches for detecting low abundance proteins (i.e., high capacity zoom gels, immunoenrichment, affinity enrichment, fluorescence dyes, etc.) are in the process of being developed and improved.

### Developing biomarkers

It was agreed that biomarkers should develop naturally from the NCT effort in toxicogenomics. However, several issues were mentioned that may inhibit development of clinical biomarkers. For example, technology transfer issues are complex and difficult to resolve quickly. In addition, these studies are difficult because environmental exposure is usually the result of exposure to an unknown substance at an unknown dose. To be useful, biomarkers must be of high specificity and sensitivity. Biomarkers of exposure and effect will be equally important and desirable. Once developed, biomarkers must be carefully validated.

### Database development

Many discussants expressed support for the proposed NCT database. One important benefit of the database would be to provide a catalogue of the results of completed experiments, thus preventing unnecessary duplication of effort. This is important because of the time- and cost-intensive nature of toxicogenomics research.

### NCT collaborations and resources

Several discussants underlined the importance of finding the right researchers for the NCT initiative. They felt that the success of the NCT depends on finding talented, committed researchers who are at the forefront of their research areas. In addition, it may be important to let researchers define their own research areas and choose their own collaborators. Work within a Consortium framework was encouraged. NIEHS Centers were also mentioned as a good model of how to promote interdisciplinary research collaborations. Discussants agreed that NCT will need to provide funds to researchers to promote toxicogenomics research. Several mechanisms were mentioned including building infrastructure, new technology grants (i.e., R21 and R33) or direct cost negotiations.

### Personnel/Training

Concern was expressed over the lack of appropriately trained scientists especially in bioinformatics. In addition, the cost of retaining good scientists in bioinformatics is often underestimated. Several discussants suggested that NCT make a concerted effort to promote training programs in bioinformatics and other fields important to the NCT effort. Training programs need sustained support and commitment, as well as adequate and sustained funding. It was also mentioned that NCT may need expertise in biostatistics as well as in bioinformatics.

**Data management, data sharing**

The problems associated with handling data from different platforms were mentioned. Although there was no strong opinion concerning the magnitude of this problem or how to overcome it, it was suggested that NCT could begin a pilot study to evaluate this problem.

*- Participants of the Roundtable Discussion were as follows:

Julio Celis (Danish Centre for Human Genome Research, Aarhus University), John Yates (Scripps Research Institute), Katheryn Resing (Howard Hughes Medical Institute, University of Colorado), Stanley Hefta (Briston Myers Squibb), Dan Liebler (Southwest Environmental Health Sciences Center, University of Arizona), Robert Wells (Texas A&M), Serrine Lau (NIEHS Center at University of Texas MD Anderson and University of Texas, Austin), Thomas Baldwin (University of Arizona), Vicki Wysocki (University of Arizona), Dave Hill (Dana Farber Cancer Center), Richard Caprioli (Vanderbilt University), Ralph Bradshaw (University of California, Irvine), Cindy Afshari (NIEHS), Allen Dearry (NIEHS), Alex Merrick (NIEHS), Kenneth Tomer (NIEHS)